***Klebsiella* MALDI TypeR: a web-based tool for *Klebsiella*
identification based on MALDI-TOF mass spectrometry**

Running title: Web-based MALDI-TOF identification of *Klebsiella*

Sébastien Bridel¹, Stephen C. Watts², Louise M. Judd², Taylor Harshegyi², Virginie Passet¹,
Carla Rodrigues¹, Kathryn E. Holt^{2,3}, and Sylvain Brisse^{1*}

¹ Biodiversity and Epidemiology of Bacterial Pathogens, Institut Pasteur, Paris, France

² Department of Infectious Diseases, Central Clinical School, Monash University, Melbourne,
Victoria 3004, Australia

³ London School of Hygiene & Tropical Medicine, London WC1E 7HT, UK

*** Corresponding author:** S. Brisse. Biodiversity and Epidemiology of Bacterial Pathogens,
Institut Pasteur, 25 rue du Docteur Roux, F-75724 Paris, France. E-mail:
sylvain.brisse@pasteur.fr; Phone +33 1 45 68 83 34

Abstract

Klebsiella pathogens affect human and animal health and are widely distributed in the environment. Among these, the *Klebsiella pneumoniae* species complex, which includes seven phylogroups, is an important cause of community and hospital infections. The *Klebsiella oxytoca* species complex also causes hospital infections and antibiotic-associated haemorrhagic colitis. The unsuitability of currently used clinical microbiology methods to distinguish species within each of these species complexes leads to high rates of misidentifications that are masking the true clinical significance and potential epidemiological specificities of individual species.

We developed a web-based tool, *Klebsiella* MALDI TypeR, a platform-independent and user-friendly application that enables uploading MALDI-TOF mass spectrometry data in order to identify *Klebsiella* isolates at the species complex and phylogroup levels. The tool, available at <https://maldityper.pasteur.fr>, leverages a database of previously identified biomarkers that are specific for species complexes, individual phylogroups, or related phylogroups. We obtained 84% to 100% identification accuracy depending on phylogroup. Identification results are obtained in a few seconds from batches of uploaded spectral data.

Klebsiella MALDI TypeR enables fast and reliable identification of *Klebsiella* strains that are often misidentified with standard microbiological methods. This web-based identification tool may be extended in the future to other human bacterial pathogens.

Keywords

36 *Klebsiella*, mass spectrometry, software

Introduction

Bacteria of the genus *Klebsiella* are human and animal pathogens with a wide ecological distribution [1]. The genus *Klebsiella*, a member of the family *Enterobacteriaceae*, encompasses a huge diversity in terms of phylogenetic lineages, genomic content, pathogenic properties, and ecological distribution. Two major subdivisions of *Klebsiella* are the *Klebsiella pneumoniae* species complex (KpSC) and the *Klebsiella oxytoca* species complex (KoSC) [2]. The KpSC currently includes five taxonomic species, two of which comprise two subspecies. These taxa were initially defined as seven distinct phylogroups, Kp1 to Kp7 [3]. *K. pneumoniae* (phylogroup Kp1) is one of the most problematic pathogens associated with antibiotic resistance worldwide. This species is phylogenetically closely related to *K. quasipneumoniae* [subsp. *quasipneumoniae* (Kp2) and subsp. *similipneumoniae* (Kp4)], *K. variicola* [subsp. *variicola* (Kp3) and subsp. *tropica* (Kp5)], *K. quasivariicola* (Kp6) and *K. africana* (Kp7). In turn, the KoSC comprises 6 species and 7 phylogroups [4]: *K. michiganensis* (Ko1), *K. oxytoca* (phylogroup Ko2), *K. spallanzanii* (Ko3), *K. pasteurii* (Ko4), *K. grimontii* (Ko6) and *K. huaxiensis* (Ko8). *K. oxytoca* is well known as a cause of opportunistic human infections and post-antibiotic hemorrhagic diarrhea [4,5].

Recent literature has brought into focus the difficulties that microbiology laboratories encounter in the identification of individual members of the *K. pneumoniae* and *K. oxytoca* species complexes [5–9]. Indeed, *Klebsiella* species are genetically similar and have few biochemical discriminatory biomarkers. As a consequence, the clinical significance of individual *Klebsiella* species and their potential epidemiological specificities are still not well defined. While Matrix Assisted Laser Desorption Ionization - Time of Flight (MALDI-TOF)

59 mass spectrometry can identify *Klebsiella* species through peak discrimination [4,5,10], the
60 current software implementation is limited to identifying *K. pneumoniae*, *K. variicola* and *K.*
61 *oxytoca* [9–11]. Here we present *Klebsiella* MALDI TypeR, a tool that leverages novel
62 discriminant biomarkers in MALDI-TOF spectra to expand the number of identifiable
63 *Klebsiella* species from MALDI-TOF data and improve identification accuracy, using a
64 previously-published method [12].

Material and methods

Bacterial strains and reference identification by genomic sequencing. A total of 341 strains were included. A first set consisted of 78 clinical isolates (50 belonging to the *Klebsiella pneumoniae* species complex, 28 to the *Klebsiella oxytoca* species complex), selected to be representative of the phylogenetic diversity of these taxa. These isolates have been used in previous studies [4,5]. Three and 24 spectra per isolate were recorded for the KpSC and for the KoSC, respectively, as described previously [4,5]. A second set of 262 isolates belonging to the KpSC was collected in Australia from humans and animals. Isolate genomes were sequenced using Illumina paired-end technology and assembled using SPAdes. Species identification was performed with Kleborate [13,14], which compares each genome to a curated set of species references using Mash distances; these species calls were used as reference in subsequent analyses.

Preparation of bacterial samples and MALDI-TOF MS analysis. An overnight culture on Luria-Bertani agar (37°C, 18 h) was used to prepare the samples with the ethanol/formic acid extraction procedure following the manufacturer recommendations (Bruker Daltonics, Bremen, Germany). Samples (1 µL) were spotted onto an MBT Biotarget 96 target plate, air dried and overlaid with 1 µL of a saturated α -cyano-4-hydroxycinnamic acid matrix solution in 50% of acetonitrile and 2.5% of trifluoroacetic acid. External calibration of the mass spectra was performed using Bruker Bacterial Test Standard. For the assessment of the direct colony picking protocol, a single bacterial colony was picked with a sterile toothpick and directly deposited onto a 96-spot polished steel target and processed as the sample spot previously mentioned. In addition, the same two protocols were followed

using blood agar medium (Columbia agar plus 5% sheep blood). A MALDI Biotyper Microflex LT controlled by Compass flexControl software (version 3.4; Bruker Daltonics, Billerica, Massachusetts, USA) was used to generate mass spectra for all isolates. Mass spectra were acquired using automatic mode and default settings (2000 to 20 000 Da; linear positive mode; 240 laser shots).

Identification of phylogroups based on MALDI-TOF data in *Klebsiella* MALDI TypeR. The KpSC and the KoSC encompass species and subspecies that are defined as phylogroups. As such, we aimed to distinguish them on the basis of MALDI-TOF data, using information from the spectra, defined as peak shifts [15]. Identification of phylogroups is therefore based on a small selection of biomarkers for which one (or more) protein weight modifications have been previously reported [4,5]. In fact, each phylogroup could be seen as comprising MALDI-Types, as described by Bridel *et al.* 2020 [12]: a specific combination of protein m/z values of these biomarkers. *Klebsiella* MALDI TypeR was run on a 4 CPUs 16-GB RAM CentOS Linux 7 (Core) virtual machine.

Results and Discussion

Development of *Klebsiella* MALDI TypeR and its web interface. *Klebsiella* MALDI TypeR was developed as a web application with R (version 3.4.0) and R Shiny (version 1.4). A user can upload either a single spectrum or multiple spectra (a maximum number of spectra is not defined, but we recommend working with batches of up to 100 spectra per analysis). The uploaded spectra must be in zip archive format. The zip archive must contain the raw data from any spectrometer. This includes the tree structure, which must

include spectra in the expected format (txt, csv, fid, ciphergen, mzXML, mzML, imzML, analyze, cdf or msd – see MALDIquant Foreign documentation for further details). This application has been tested with spectra (raw data folders) from Bruker spectrometer only but could theoretically handle any proper archive containing one of the previously mentioned file formats. Supplementary Figure 1 describes each step to reach one raw data folder. Each archive may correspond either to a single raw spectrum or several raw spectra of the same isolate (technical replicates). For better results, we recommend using at least 3 technical replicates for each isolate.

The spectra can be visually inspected after upload (**Figure 1**). A panel shows each uploaded spectrum for quick visual inspection (*e.g.*, intensity checking) and a table highlights the presence (if any) of empty spectra.

For identification, the tool was designed with a 2-step process. The first step (**Figure 2, panel B1**) identifies isolates at the species complex level using two sets of biomarkers, each set corresponding to reference biomarkers for (i) the KpSC (29 biomarkers, **Table S1**) or (ii) the KoSC (57 biomarkers, **Table S1**). This step can be performed, based on user choice, with one of two alternative peak detection algorithms. The first one is based on the R package MALDIquant [16] and is set by default, as it runs faster. The second one uses the MassSpecWavelet package [17] and generally detects more significant peaks, but is slower. Species complex identification is based on the ratio of common peaks between the sample peak list and a reference peak list, to the number of peaks in the corresponding reference (this ratio is called “Identification Value”). The input sample is assigned to the species complex with the highest Identification Value.

The second step consists in the identification of the specific phylogroup (corresponding to a species or subspecies) within the defined species complex (**Figure 2, panel B2**). The raw spectrum is processed again by the Multi Peak Shift Typing (MPST) method [12], but this time using the set of biomarkers corresponding to the assigned species complex, which allows recalibration of m/z values. The biomarkers used for the two MPST schemes (one for each species complex, see **Tables 1 to 4**) are based on previous work [4,5]. Briefly, an MPST scheme is analogous to a multi-locus sequence typing (MLST) scheme where an allele type (AT) corresponds to an isoform (IF) for one protein (each IF is defined by a specific mass – or a m/z value on the spectrum) and a MALDI-Type (MT) corresponds to a sequence type (ST). Each MT was observed within a single phylogroup. The tool attempts to define the IF for each biomarker, and then to attribute the spectrum to the phylogroup corresponding to the relevant MT (**Figure 2, panel B3**). The only exception to this process was introduced for the discrimination of *K. oxytoca* phylogroups Ko1, Ko4 and Ko6 (*K. oxytoca*, *K. pasteurii* and *K. grimontii*, respectively). These three species form a phylogenetic branch of closely related species within the *K. oxytoca* species complex [4], and their spectra are very similar such that they share the same MT. In order to distinguish among these three phylogroups, we added 3 previously described biomarkers [4], for which the tool checks for presence/absence within the spectra (*i.e.*, these 3 biomarkers do not correspond to peak shifts). If any of the above biomarkers (peak shifts or presence/absence) is not found, automatic identification is not considered possible, and the result is reported to the user as NT (not typable). In this case, manual analysis or data re-acquisition are needed.

To help resolve these cases and allow users to further explore the data, we included a feature enabling to interactively explore the spectra. The plot window (**Figure 3**) shows the

sample spectra overlaid on the reference spectra (1 to 3 spectra for each phylogroup).
Predefined windows (one for each biomarker) allow the user to directly investigate specific
spectral regions. This feature could also enable the detection of new isoforms or biomarkers.
It also contains a customizable option that allows the user to select any region of the spectra.

The *Klebsiella* MALDI TypeR tool was made available at <https://maldityper.pasteur.fr>.
A single isolate analysis required approximately 10 seconds to complete and the analysis of
10 isolates takes about 35 seconds. The tool does not have a limit to the number of isolates
that can be analyzed simultaneously, and it performed well with batches of 100 isolates that
were uploaded at once.

Evaluation of the identification tool. We first tested fifty KpSC isolates grown
overnight on lysogeny broth (LB) agar (37°C, 18h) and subjected to ethanol/formic acid
extraction following the manufacturer recommendations (Bruker Daltonics, Bremen,
Germany). For each strain, a total of 3 spectra from the same spot were acquired (Bruker
Microflex LT mass spectrometer, Paris). This dataset yielded correct assignments to species
complex for 50/50 (100%) isolates, 47 (94%) of which had an identification value >50%
(**Tables S2 and S3**). The same samples prepared using different experimental conditions (two
media: LB and Columbia agar plus 5% sheep blood; each with both extraction and direct
application in the MALDI-TOF target) also yielded correct complex assignments for all
isolates, albeit with lower confidence (**Table S2**). Regarding phylogroup assignment, using
LB-extraction 90% of samples were correctly identified at phylogroup level (**Table S3**), while
the other sample preparation methods yielded better results overall (**Table S4**). To validate
robustness of the tool to instrument and user variation, we tested an additional dataset

comprising 248 KpSC isolates prepared using the LB-extraction method (176 isolates with a single spectrum, 26 in duplicate and 68 in triplicate); these spectra were generated on a different instrument (Bruker MALDI Biotyper microflex LT-SH) in a different laboratory (Melbourne, Australia). 244/248 isolates (98%) were assigned to the correct species complex, and 219/244 (90%) were correctly identified at the phylogroup level (84% to 100% for each phylogroup, see **Table S5**).

The KoSC assignments were tested using 28 isolates prepared using LB-extraction and triplicate spectra generated on the Bruker Microflex LT mass spectrometer instrument (Paris). All isolates were correctly assigned to complex level and most (89%) were assigned to the correct MALDI-Type (**Tables S6 and S7**). However, discrimination of MALDI-Type MT1 (n=16, 57%) into phylogroups Ko1, Ko4 and Ko6 was poor (**Table S8**) because the three additional biomarkers intended to discriminate between these phylogroups were captured only rarely.

Conclusions

We developed the *Klebsiella* MALDI TypeR web tool for identification of *Klebsiella* species based on MALDI-TOF MS. Our tool allows the fast analysis of single or multiple spectra, including their visualization. We show that MALDI-TOF MS leads to accurate identification for *Klebsiella* phylogroups of the KpSC, the major *Klebsiella* species complex found in clinical samples. *Klebsiella* MALDI TypeR therefore represents an important resource for the identification of the KpSC members, which are currently difficult to identify in routine microbiology practice. For the KoSC, we found promising data, but further

validation based on a larger dataset would be required. The current biomarker approach does not appear to be suitable for accurate discrimination of Ko1, Ko4 and Ko6. In the future, the website may be updated to integrate additional methods for identification and typing, such as supervised learning.

Declarations

Ethics approval and consent to participate

Not applicable

Consent for publication

Not applicable

Availability of data and materials

The datasets used and/or analyzed during the current study are available from the corresponding author on reasonable request.

Competing interests

The authors declare that they have no competing interests.

Funding

This work was financially supported by the MedVetKlebs project, a component of European Joint Programme One Health EJP, which has received funding from the European Union's Horizon 2020 Research and Innovation Programme under Grant Agreement No. 773830.

218 **Authors' contributions**

219 LMJ, TH, CR and VP performed the MALDI-TOF data acquisition. SBridel, SCW and CR
220 analyzed the MALDI-TOF data. SBridel designed and programmed the initial version of the
221 *Klebsiella* MALDI TypeR tool and website. SCW, CR, KEH and SBrisse provided input in
222 the choice of data processing algorithms and web site design. SBridel wrote the original draft.
223 All authors contributed to and approved the final version of the manuscript.

224

225 **Acknowledgements**

226 We acknowledge the help of Stevonn Volant and Amine Ghoulane from the Bioinformatics
227 and Biostatistics HUB of Institut Pasteur for help with R Shiny deployment of the web tool on
228 the web server.

229 References

- 230 [1] Brisse S, Grimont F, Grimont PAD. The genus *Klebsiella*. In: Dworkin M, Falkow S,
 231 Rosenberg E, Schleifer K-H, Stackebrandt E, editors. *Prokaryotes · Handb. Biol. Bact.*, vol. 6:
 232 *Proteobacteria: Gamma Subclass*. 3rd edition, New York: Springer; 2006.
- 233 [2] Wyres KL, Lam MMC, Holt KE. Population genomics of *Klebsiella pneumoniae*. *Nat*
 234 *Rev Microbiol* 2020;18:344–59. <https://doi.org/10.1038/s41579-019-0315-1>.
- 235 [3] Rodrigues C, Passet V, Rakotondrasoa A, Diallo TA, Criscuolo A, Brisse S.
 236 Description of *Klebsiella africanensis* sp. nov., *Klebsiella variicola* subsp. *tropicalensis* subsp.
 237 nov. and *Klebsiella variicola* subsp. *variicola* subsp. nov. *Res Microbiol* 2019;170:165–70.
 238 <https://doi.org/10.1016/j.resmic.2019.02.003>.
- 239 [4] Merla C, Rodrigues C, Passet V, Corbella M, Thorpe HA, Kallonen TVS, et al.
 240 Description of *Klebsiella spallanzanii* sp. nov. and of *Klebsiella pasteurii* sp. nov. *Front*
 241 *Microbiol* 2019;10. <https://doi.org/10.3389/fmicb.2019.02360>.
- 242 [5] Rodrigues C, Passet V, Rakotondrasoa A, Brisse S. Identification of *Klebsiella*
 243 *pneumoniae*, *Klebsiella quasipneumoniae*, *Klebsiella variicola* and Related Phylogroups by
 244 MALDI-TOF Mass Spectrometry. *Front Microbiol* 2018;9.
 245 <https://doi.org/10.3389/fmicb.2018.03000>.
- 246 [6] Brisse S, van Himbergen T, Kusters K, Verhoef J. Development of a rapid
 247 identification method for *Klebsiella pneumoniae* phylogenetic groups and analysis of 420
 248 clinical isolates. *Clin Microbiol Infect* 2004;10:942–5. <https://doi.org/10.1111/j.1469-0691.2004.00973.x>.
- 250 [7] Seki M, Gotoh K, Nakamura S, Akeda Y, Yoshii T, Miyaguchi S, et al. Fatal sepsis
 251 caused by an unusual *Klebsiella* species that was misidentified by an automated identification
 252 system. *J Med Microbiol* 2013;62:801–3. <https://doi.org/10.1099/jmm.0.051334-0>.
- 253 [8] Fonseca EL, Ramos N da V, Andrade BGN, Morais LLCS, Marin MFA, Vicente
 254 ACP. A one-step multiplex PCR to identify *Klebsiella pneumoniae*, *Klebsiella variicola*, and
 255 *Klebsiella quasipneumoniae* in the clinical routine. *Diagn Microbiol Infect Dis* 2017;87:315–
 256 7. <https://doi.org/10.1016/j.diagmicrobio.2017.01.005>.
- 257 [9] Long SW, Linson SE, Saavedra MO, Cantu C, Davis JJ, Brettin T, et al. Whole-
 258 Genome Sequencing of Human Clinical *Klebsiella pneumoniae* Isolates Reveals
 259 Misidentification and Misunderstandings of *Klebsiella pneumoniae*, *Klebsiella variicola*, and
 260 *Klebsiella quasipneumoniae*. *MSphere* 2017;2. <https://doi.org/10.1128/mSphereDirect.00290-17>.
- 262 [10] Dinkelacker AG, Vogt S, Oberhettinger P, Mauder N, Rau J, Kostrzewa M, et al.
 263 Typing and Species Identification of Clinical *Klebsiella* Isolates by Fourier Transform
 264 Infrared Spectroscopy and Matrix-Assisted Laser Desorption Ionization–Time of Flight Mass
 265 Spectrometry. *J Clin Microbiol* 2018;56. <https://doi.org/10.1128/JCM.00843-18>.
- 266 [11] Berry GJ, Loeffelholz MJ, Williams-Bouyer N. An Investigation into Laboratory
 267 Misidentification of a Bloodstream *Klebsiella variicola* Infection. *J Clin Microbiol*
 268 2015;53:2793–4. <https://doi.org/10.1128/JCM.00841-15>.
- 269 [12] Bridel S, Bourgeon F, Marie A, Saulnier D, Pasek S, Nicolas P, et al. Genetic diversity
 270 and population structure of *Tenacibaculum maritimum*, a serious bacterial pathogen of marine
 271 fish: from genome comparisons to high throughput MALDI-TOF typing. *Vet Res* 2020;51:60.

272 <https://doi.org/10.1186/s13567-020-00782-0>.
 273 [13] Holt K. katholt/Kleborate. 2020.
 274 [14] Lam MMC, Wick RR, Wyres KL, Holt KE. Genomic surveillance framework and
 275 global population structure for *Klebsiella pneumoniae*. *BioRxiv* 2020:2020.12.14.422303.
 276 <https://doi.org/10.1101/2020.12.14.422303>.
 277 [15] Sauget M, Valot B, Bertrand X, Hocquet D. Can MALDI-TOF Mass Spectrometry
 278 Reasonably Type Bacteria? *Trends Microbiol* 2017;25:447–55.
 279 <https://doi.org/10.1016/j.tim.2016.12.006>.
 280 [16] Gibb S, Strimmer K. MALDIquant: a versatile R package for the analysis of mass
 281 spectrometry data. *Bioinformatics* 2012;28:2270–1.
 282 <https://doi.org/10.1093/bioinformatics/bts447>.
 283 [17] Du P, Kibbe WA, Lin SM. Improved peak detection in mass spectrum by
 284 incorporating continuous wavelet transform-based pattern matching. *Bioinforma Oxf Engl*
 285 2006;22:2059–65. <https://doi.org/10.1093/bioinformatics/btl355>.

286